DOCUMENT RESUME

ED 463 319 TM 033 751

AUTHOR Li, Yuan H.
TITLE An Evaluation of the Construct Validity for the Multiple-Subject Testing Programs.
PUB DATE 2001-04-00
NOTE 29p.; Paper presented at the Annual Meeting of the American Educational Research Association (Seattle, WA, April 10-14, 2001).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Achievement Tests; *Construct Validity; Correlation; Elementary Education; *Elementary School Students; Performance Based Assessment; State Programs; Structural Equation Models; *Testing Programs; True Scores
IDENTIFIERS *Comprehensive Tests of Basic Skills; *Maryland School Performance Assessment Program

ABSTRACT

The primary objective of this study was to examine the construct validity of two multiple-content testing programs, the multiple-choice Comprehensive Tests of Basic Skills (CTBS/5) and the performance-based Maryland School Performance Assessment Program (MSPAP), by evaluating the true-score longitudinal associations among multiple-content scores in one school. The CTBS dataset contained scores for 6,841 students with 5 content area scores on 2 grade-level tests. The first set for the MSPAP analyses contained scores for 6,326 students with 6 content area scores on 2 grade-level tests. For cross validation, the second dataset contained 6,547 students. Whether the true-score correlation between two time-period measures of the same content area was higher than the true score correlations with other content areas was evaluated. This criterion was achieved in two (Reading and Mathematics) of the five CTBS/5 content subtests, as well as one (Language) of the six MSPAP content subtests. The structural equation modeling was conducted with a multitrait-multimethod correlation dataset. The traits of Reading and Mathematics were assessed by MSPAP and the old version of the CTBS/4. Although convergent validity existed for these two measures, there was little evidence to support discriminant validity for both measures. (Contains 7 tables and 19 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made from the original document.

An Evaluation of the Construct Validity for the Multiple-subject Testing Programs

by

Yuan H. Li
Prince George's County Public Schools

ED 463 319

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Y. Huang Li

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

1

Paper was presented at the annual meeting of the American Educational Research Association, April, 10-14, 2001, Seattle, WA.

TM033751

BEST COPY AVAILABLE

Abstract

The primary objective of this study was to examine the construct validity for the two multiple-content testing programs, the multiple-choice Comprehensive Tests of Basic Skills (CTBS/5) together with the performance-based Maryland School Performance Assessment Program (MSPAP), by evaluating the true-score longitudinal associations among multiple-contest scores in one school district.

The following criterion was closely examined: the true-score correlation between two time-period measures of the same content area is higher than its longitudinal true-score correlations with other content areas. This criterion was achieved in two (Reading and Mathematics) of five CTBS/5 content subtests, as well as one (Language) of six MSPAP content subtests.

The structural equation modeling has been conducted on a multitrait-multimethod correlation dataset, where the traits of Reading and Mathematics were assessed by MSPAP and the old version of CTBS/4. Although convergent validity existed in these two measures, there was little evidence to support discriminant validity in both measures.

Key Words: Construct Validity, Performance Assessment, Multiple-choice Assessment, Longitudinal Correlation, Structural Equation Modeling (SEM).

I. Introduction

This study examined the construct validity of two multiple-content assessment programs that are part of the overall accountability and school improvement initiative in the State of Maryland. Construct validity has been the subject of considerable discussion and debate historically in test literature, especially in recent years due to the use of performance assessment and high stakes performance accountability decisions. This study examined results from one of the larger school districts in the State of Maryland. The two assessment measures are briefly introduced below.

A. Background of the Two Multiple-content Testing Programs

MSPAP

The first achievement measure is the Maryland School Performance Assessment Program (MSPAP, Maryland State Department of Education, 1998), a unique performance-based assessment initiated in the 1990-91 school year. The MSPAP was administered to students in grades 3, 5, and 8 in all of its public schools. It consists of six content areas: Reading, Writing, Language Usage, Mathematics, Science and Social Studies. MSPAP test items (tasks) are integrated both within a content area and across content areas so that students have an opportunity to assimilate information they have learned. To cover the required breadth of learning outcomes in limited testing time, three non-parallel test forms per content area were developed and randomly assigned to students within a school. Because of the design of the MSPAP test and its sampling design, the primary focus of the information provided from MSPAP assessments is school performance, rather than individual student performance.

CTBS/5 or CTBS/4

The second assessment measure, the Comprehensive Test of Basic Skills (CTBS/5, Survey, CTBS/McGraw-Hill, 1997), was administered statewide as a school accountability index for the first time since 1999. The CTBS/5 was administered for all second, fourth and sixth grade students. The CTBS/5 is the multiple-choice assessment that consists of five content areas:

Reading Expression, Language Usage, Language Mechanics, Mathematics Concepts and Mathematics Computation.

Under the current two assessment programs, no students are allowed to take both MSPAP and CTBS/5 tests in the same school year. However, part of the research design (discussed later) for this study requires a dataset, in which students have both MSPAP and CTBS scores from the same school year. In the previous school years 1995 and 1996, the old version of CTBS/4 and MSPAP were administered to the third grade students at the same school year in both Fall and Spring, respectively. The MSPAP/CTBS4 test data collected in school year 1995 was selected to serve this purpose. The CTBS/4 consists of five content areas: Reading Vocabulary, Reading Comprehension, Spelling, Math Computation, and Math Applications.

The newest version of CTBS/5 was designed to provide scale-score continuity with the previous version of CTBS/4 to facilitate evaluation of instructional effectiveness and performance growth over a period of time. Hence, for the same content subtest between the two tests, a look-up table with equivalent scores for the two tests has been constructed using the equpercentile equating procedure (for the illustrations of test equating, see Kolen, 1995). The CTB test publishing company provides this type of information to its test users. Overall, the statistical characteristics (e.g., test difficulty and test reliability) are similar for the two tests; however, assessment contents of the CTBS/5 are more integrated to reflect current curricula and classroom practices (for more detailed comparisons between CTBS/5 and CTBS/4, refer to the technical report, CTB McGraw-Hill, 2001).

B. Construct Validity of MSPAP/CTBS Assessments

MTMM

The multitrait-multimethod (MTMM) developed by Campbell and Fiske (1959) was often used for examining the construct validity of the multiple-content measures such as MSPAP and CTBS. The MTMM model includes four types of correlation (Nunnally & Bernstein, 1994). They are: (1). The correlation (reliability coefficient) between the same trait scores measured by the same assessment methods, (2). The correlation (convergent validity coefficient) between the same trait scores measured by different assessment methods, (3).The correlation (discriminant

validity coefficient, called heterotrait-monomethod coefficient) between two different trait scores measured by the same measurement methods, and (4). The correlation (discriminant validity coefficient, called heterotrait-heteromethod coefficient) between two different trait scores measured by different assessment methods. For a content measure, if its coefficients of reliability, convergent validity, heterotrait-monomethod and heterotrait-heteromethod are in the order from largest to smallest, the evidence of construct validity for this content measure is presumed established.

Schatz (1998) applied the MTMM approach to examine the reliability-convergent validity coefficients for the MSPAP/CTBS4 reading and mathematics achievement scores. In that study, reading and mathematics were assessed by the MSPAP performance-based assessment and by the two multiple-choice measures, CTBS/4 together with a Criterion Referenced Test. The expected order of correlation coefficients, indicated in the above MTMM model, was found for the content area of Mathematics at three grade levels, Grades 3, 5 and 8, but the content area of Reading did not fit the expected pattern at any of the three grade levels. Was this problem caused by the performance-based assessment or by the multiple-choice assessment? The answer to this question based on the analysis of MTMM correlation was unclear. In addition, visual inspection for assessment of construct validity data in a correlation matrix can be problematic because of measurement and sampling errors.

SEM /MTMM

When the structural equation modeling (SEM) is applied to the data collected from the MTMM method (for literature review, see Schmitt & Stults, 1986), this SEM's application may relieve part of the problems that MTMM has encountered. The SEM/MTMM is capable of testing the convergent and discriminant validities. Also, it can furthermore partition the variance of each content measure into three components: specific trait, assessment method, and random error. The comparisons among the magnitudes of the three components for each content measure are additional for evaluating the construct validity of a testing program. Li, Ford and Tompkins (1999) employed SEM on the test data collected by the multitrait-multimethod, where the traits of Reading and Mathematics were assessed by MSPAP together with CTBS/4. Their results demonstrated that despite evidence of convergent validity for these two measures, assessment method effects were instrumental in attenuating trait variances.

Modeling a SEM model to a set of data has its limitations. The results yielded from a SEM modeling are based solely on the set of data for which the fit is optimized. SEM might result in different findings when data differ. What is needed for fitting a SEM to a MTMM data is to reexamine how well this SEM model will hold up for future data. This process is called cross-validation (Tatsuoka & Lohnes, 1988) and has been incorporated into the design of this study.

Evaluating the Longitudinal True-score Association

An alternative of examining the construct validity of multiple-content measures is to evaluate the longitudinal true-score association for the matched-sample test data across two different time period measures. For instance (see Li et al., 1999) test scores were collected for students who had multiple-content scores on two MSPAP measures when they were in the third grade in 1994 and in the fifth grade in 1996. The longitudinal true-score correlation between the two same content measures across two time periods (e.g., two years) was then computed (for detailed computing procedures, see the next section). This longitudinal true-score correlation is not expected to be very high because of examinees' maturational and changing test specifications between testings. How high is significantly enough? Reviewing the literature, there is no reference addressing this issue. Li et al. (1999) adopted a relative rather than absolute criterion to deal with this issue. That is: the true-score correlation between two time-period measures of the same content area is higher than its true-score longitudinal correlations with other content areas. Once this criterion is achieved, some evidence of construct validity is presumed found in the contest subtest being examined. The underlying principles in regard to this idea are illustrated later.

In Li et al.'s study (1999), all MSPAP content area measures did not meet this criterion. Because of that result, the need for closely examining more longitudinal MSPAP test data becomes necessary. Also, evaluating the longitudinal true-score associations for the other assessment program, CTBS/5, is of interest and critical.

C. Study Purposes

In a school improvement instructional model that includes high-stake testing programs (Linn, 1995), data resulting from both the performance-based and the multiple-choice assessments must provide school managers and teachers with valid information for assessing the strengths and weaknesses in their instructional programs. In order to achieve this goal, examining the construct validity of the two testing programs is a primary aim of this study.

The method of evaluating the longitudinal true-score association is one of the primary methods for serving the purpose of this study. Two sets of longitudinal CTBS/5 (or MSPAP) test data (illustrated later) were collected and then evaluated for the construct validity for each content area measure. The results from one set of longitudinal data were then compared with those from another set. This cross-validation procedure makes findings from this study more convincing and reliable.

As noted earlier, SEM modeling is very sensitive to the data. Hence, the method of SEM/MTMM was also employed to reexamine whether the SEM models used in Li et al.' study (1999) would hold up for a different but similar dataset (illustrated later).

Hopefully, findings from both evaluating methods will provide valuable information of construct validity for these two multiple-content testing programs.

II. Overview of Statistical Procedures

Several methods exist for examining the validity of a testing program. No precise method can be used to draw a final conclusion as to any testing program. The results gathered from multiple methods would help researchers reach a more accurate conclusion than those from only a single method. This section briefly reviews two methods used in this study. For readers interested in this field, several valuable references (e.g., Byrne & Bazana, 1996; Crocker & Algina, 1986; Nunnally & Bernstein, 1994; Schmitt, 1986) are available.

A. Method of Testing the Convergent and Discriminat Validities

How do we know that the characteristics of the convergent and discriminat validities exist in the MTMM data of interest? The Widaman's paradigm (1985) serves this purpose. When applying his approach, four specific SEM models are created as shown below, using a test data from the CTBS/4 and MSPAP measures as an example. It should be noted that the four models shown below are by no means exhaustive. Other alternative models may be applicable.

Suppose, the SEM is used to model the MTMM data for the five content area scores. They consist of MSPAP Reading (MSPAPRD), MSPAP Math (MSPAPMS), CTBS Reading Vocabulary (CTBRVS) , CTBS Reading Comprehension (CTBSRCS) and CTBS Math Applications (CTBSMAS). The relationships of these five content measures with the latent factors are illustrated below.

For a SEM path diagram, observed variables are shown in boxes and latent factors in ellipses (or circles). In reviewing a SEM model depicted in Figure 1, there are five observed variables and four latent factors. The two-way arrows represent covariances or correlations between pairs of variables. The unidirectional arrows leading from factors (e.g., READING) to each of the observed variables (e.g., MSPAP Reading, CTBS Reading Vocabulary) suggest that scores on the observed variables are caused by the latent factors. The sourceless one-way arrows pointed from the Es (e.g., E1) indicate the impact of random measurement error on the observed variables (e.g., MSPAP Reading). The standardized path coefficients presented in Figure 2 are part of the results of this research and will be discussed later.

Creating Four Specific Models

The first model is called Model M1 that has two latent trait factors, together with two latent assessment method factors. This model allows the two latent traits to be correlated, along with the two method effects to be correlated (see Figure 1). Specifically, it is hypothesized that the latent READING trait is measured by MSPAP Reading, CTBS Reading Vocabulary and CTBS Reading Comprehension. The observed variables of MSPAP Math along with CTBS Math Applications are hypothesized to be indicators of the latent MATH factor. It is hypothesized that the performance-based assessment (called MSPAP) has some impact on the observed variables of MSPAP Reading and MSPAP Math. Similarly, the three observed variables of CTBS Reading

Vocabulary, CTBS Reading Comprehension and CTBS Math Applications are supposed to be affected by the multiple-choice assessment (called CTBS). This model serves as the base line against which an alternative model presented below is compared. It is typically the least restrictive model.

The second model is called Model M2 in which no trait factors are specified, but the two method effects are allowed to be correlated. This model is nested within Model M1.

The third model is called Model M3 in which two traits are perfectly correlated and the two method effects are allowed to be correlated. This model is formed by fixing the correlation between two trait factors to 1.0 in the model of M1.

The fourth model is called Model M4 in which two traits are allowed to be correlated and the two method effects are perfectly correlated. This model is constructed by fixing the correlation between two method factors to 1.0 in the model of M1.

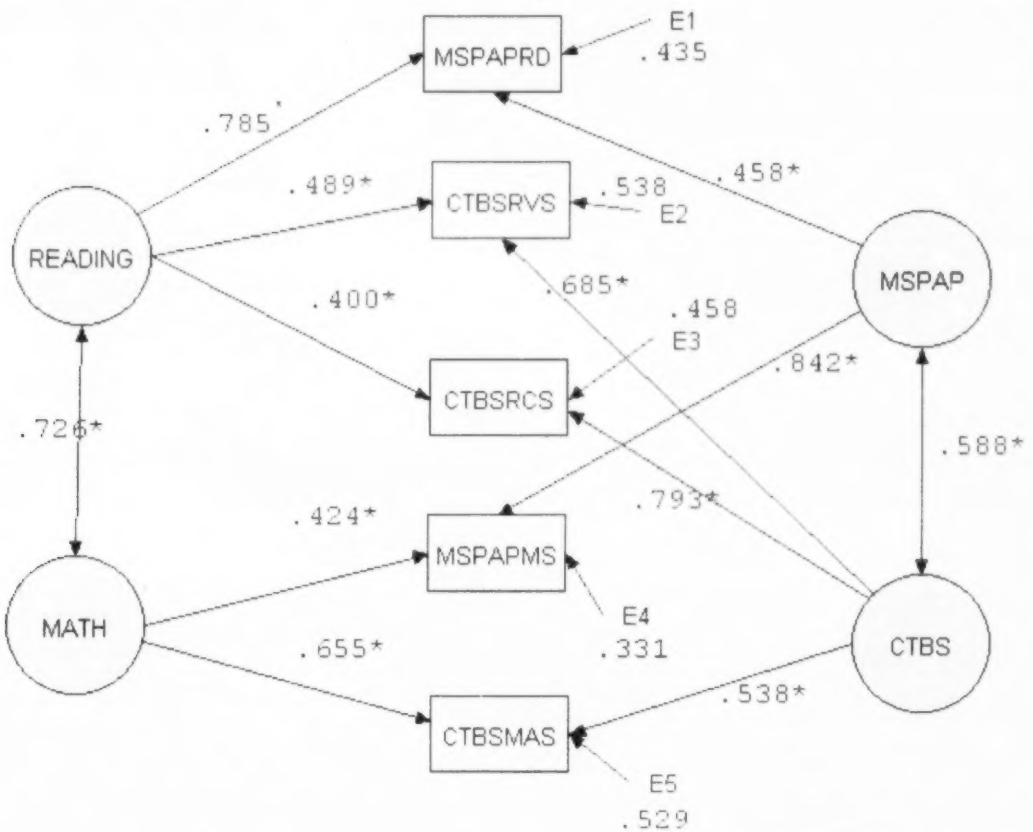

Figure 1: A Hypothesized Multitrait-multimethod Model for the MSPAP-CTBS Test Data

BEST COPY AVAILABLE

Testing Convergent Validity

Using Widaman's (1985) paradigm, the evidence of convergent validity can be tested by comparing a model in which traits are specified (Model M1) with one in which they are not (Model M2). A test of difference in chi-square values between Models 1 and 2 is conducted to test the convergent validity.

Testing Trait Discriminant Validity

In testing for evidence of discriminant validity between traits (reading and math), a comparison is made between a model in which traits correlated freely (Model M1) with one in which they are perfectly correlated (Model M3). A test of the difference in chi-square values between two models is conducted to evaluate the trait discriminant validity.

Testing Method Discriminant Validity

The same logic, as noted earlier, is used to evaluate the evidence of discriminant validity between methods (MSPAP and CTBS). A model in which method factors are freely correlated (Model M1) is compared with one in which they are perfectly correlated (Model M4). A test of the difference in chi-square values between two models is conducted to evaluate the evidence of method discriminant validity.

To summarize, hypothesis test and fit indices are used to evaluate whether models are attainable. In addition, a test in chi-square values between two nested models is used to evaluate which model is better capable of capturing the data. Finally, breaking down the variance for each observed measure into its components: specific trait, measurement method effect, and error term, was used to evaluate whether the assessment method effects attenuate the trait effect.

B. Gauging the Construct Validity for the Longitudinal Intercorrelations

Computing the True-score Longitudinal Associations

When the method of evaluating the longitudinal association is performed for a relatively large sample size, the issue of sampling error could be relatively minor. However, the

measurement errors (unreliability) of two measures can not be avoided and will cause correlation attenuation (Lord, 1980).

A correction for attenuation can be obtained by computing the true-score (without measurement error) relationship between two measures. This statistic can be estimated by either using the formula presented in Lord (1980) or using the SEM approach. When SEM is applied, the true-score correlations among multiple subtests scores can be calculated by the following procedures. A SEM diagram shown in Figure 2 is used for easier illustration. F1 and F2 in Figure 2 depicts the constructs underlying the observed variables of V1 and V2, respectively. The value of 1 (in Figure 2) next to the arrow sign represents the path coefficient, and D2 represents the impact of random error on the factor of F2.

(1). Create latent factors for each of the multiple-subtest scores with only a single measured indicator variable.

(2). Rather than estimating the error variances (e.g., E1 and E2 for the variables V1 and V2, see Figure 2), as is customary, we need to fix the error variance (e.g., E1 or E2) for each observed variable. The error variance for a subtest should be approximated by the variance of this subtest multiplied by 1-reliability coefficient of this subtest, and

(3). Compute the correlation among latent factors (e.g., F1 and F2). The values of the correlations among latent factors are the true-score correlations among multiple subtest scores. In figure 2, the value of the standardized path coefficient leading from F1 to F2 is equivalent to the value of the correlation coefficient between F1 and F2.

The above procedures can be conducted by the SEM computer softwares such as EQS (Bentler, 1995).

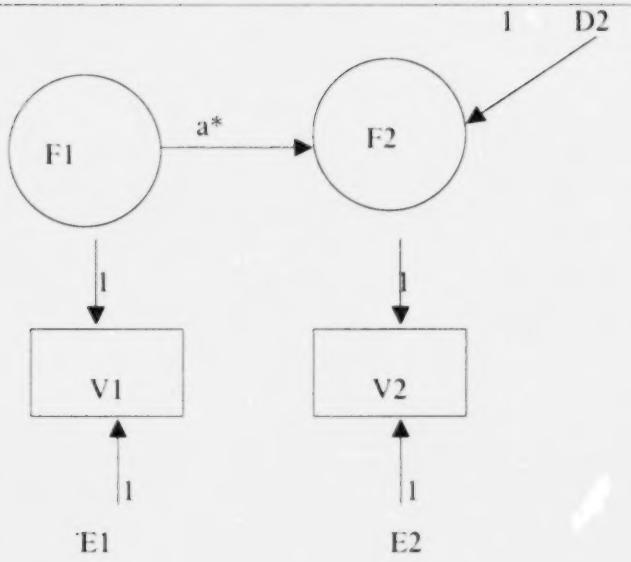

Figure 2: A Diagram to Demonstrate How to Compute the True-Score Correlation Between Two Variables

Justification of Construct Validity

1. Alternative Definition for the Longitudinal Correlation

The criterion-related validation is defined to assess the degree of relationship between a predictor and criterion (Crocker & Algona, 1986). In the context of the longitudinal test scores, the test scores obtained from the "earlier" or "later" testings can be treated as a predictor or a criterion interchangeably. Both "earlier" and "later" tests measure the same underlying latent trait in spite of their differences in test difficulty. Hence, the longitudinal correlation between two-time-period measures of the same content is a validity coefficient which gauges how well the one measure can predict the other measure or how well both measures hit the same target.

Intuitively, this validity coefficient obtained form the longitudinal correlation is like the reliability coefficient of stability or equivalence. Because both "earlier" and "later" tests are not identical and also not alternate forms, this coefficient is not the same as a reliability coefficient.

2. Gauging the Construct Validity

Once the longitudinal true-score intercorrelations among all multiple-content measures are obtained, setting the criterion for gauging the construct validity for each content measure is illustrated below.

Consider a construct validation study in which an investigator is interested in which of the three variables of X, Y, and Z is more closely related to the target construct of P variable. The investigator collects the information regarding the validity coefficients between P and these three variables, X, Y, Z. If the true-score rather than "observed-score" correlations, P with, X, Y, Z are .5, .6. 7, respectively, we will conclude that Z measure is the most closely related to the underlying construct of P variable. The above conclusion is purely based on the meaning of the true-score correlation that is calculated when random errors are removed from the two measures by using the statistical modification technique. Hence, the true-score correlation of .7 is absolutely higher than .6 or .5 when the statistical modification appropriately corrects the error variances of variables.

Suppose, before data collection, this investigator already knew that the answer to the above question should be X variable because X is specifically designed to measure the construct of P variable. However, after reviewing the validity coefficients, as shown above, this investigator might suspect the construct validity of the X measure because it's validity coefficient with P is lower than P with the variables of Y and Z that were not designed to measure the underlying construct of P.

On the other hand, if the validity coefficients, P with, X, Y, Z are replaced with .7, .6. .5 (rather than .5, .6., 7), respectively, this result suggests that X's measure to the underlying construct of P is somewhat valid to some extent because its validity coefficient with P is, at least, higher than P's validity coefficients with other variables. It appears clear that the criterion used here to gauge the existence of construct validity for a content measure (e.g., Reading) does not rely on the magnitude of its validity coefficient itself , but rather depends on whether its coefficient is higher than the others (e.g., with Science) that are irrelevant to this content measure.

Obviously, if this criterion is not achieved for a content measure, there is a lack of evidence of construct validity for this content measure. For example, the value of the

longitudinal true-score association between Reading itself is less than its true-score longitudinal association with Science, we might wonder whether the measure of the Reading construct has appropriately been assessed. On the other hand, if this criterion is met, the construct validity for this measure is presumed sound, to some extent.

The method of gauging the construct validity for a test, illustrated above, could encounter practical problems, for example, when some of the values of true-score correlation are very close, or very low themselves. As a matter of fact, the former scenario could happen in all kinds of statistical tests. For instance, when the type-I error is set to .05 to a statistical test and the type-I error is computed as .049, we should reject the null hypotheses from the perspective of the statistical test. On the other hand, people might argue as to whether the observed type-I error of .049 is practically different from its cutoff value of .050. Similar argument could occur when comparing the true-score correlation between .5 and .49. When this sort of scenario occurred, there is no doubt but that personal judgment might be involved in reaching a final conclusion for this type of result.

When all true-score longitudinal correlations are very low, we do not recommend using the criterion introduced above for gauging the construct validity for a test. If the elapsed time between two testings is not too long and the standardized procedures of two testings are properly implemented, the likelihood of this scenario happening would not be high, practically and empirically.

To summarize, the method of evaluating the longitudinal true-score association does not depend on different types of measures as MTMM does. Consequently the results obtained from this method are much easier to interpret than those from the MTMM method. In addition, as the testing programs continue to be implemented, the test data required for this method evaluation is practically easier to obtain, compared with the data required by the method of SEM/MTMM.

III. Methodology

A. Evaluating Longitudinal True-score Associations

1. Data Description and Sample Size

Matched-sample CTBS Datasets

The first dataset (CTBS1, refer to Figure 3) is the CTBS/5 test scores for 6841 students who had five content area scores on two grade-level test scores. Students took tests when they were in the second grade in Spring, 1997 and in the fourth grade in Spring, 1999. The second dataset (CTBS2 similar to the first dataset) was collected for the purpose of conducting cross validation. It included 6899 students' CTBS/5 test scores obtained when they were in the second grade in Spring,1998 and in the fourth grade in Spring, 2000.

Matched-sample MSPAP Datasets

The first dataset for the MSPAP longitudinal association analyses (MSPAP1, refer to Figure 3) comprised the test scores for 6326 students who had six content area scores on two grade-level MSPAP tests took when they were in third grade in Spring,1997 and in fifth grade in Spring, 1999.

For conducting cross-validation, the second dataset (MSPAP2) is similar to the first MSPAP dataset. It covered 6547 students' MSPAP scores obtained when they were in third grade in Spring,1998 and in fifth grade in Spring, 2000.

Dataset	School Year (SY)					
	SY95	SY96	SY97	SY98	SY99	SY00
CTBS1			Grade 2 Spring CTBS/5	→	Grade 4 Spring CTBS/5	
CTBS2				Grade 2 Spring CTBS/5	→	Grade 4 Spring CTBS/5
MSPAP1			Grade 3 Spring MSPAP	→	Grade 5 Spring MSPAP	
MSPAP2				Grade 3 Spring MSPAP	→	Grade 5 Spring MSPAP
CTMS	Grade 3 Fall CTBS/4 Spring MSPAP					

Figure 3: An Illustration for the Five Datasets Analyzed in this Study

2. Data Analysis and Evaluation

The analysis of the true-score intercorrelations among students' performance in multiple content areas across two time-period measures was performed. The criterion used for gauging the existence of construct validity for content measure was closely examined.

B. SEM/Multitrait-multimethod Associations

1. Data Description and Sample Size

In order to perform the SEM/MTMM modeling for the MSPAP and CTBS testing programs, ideal test data should contain students' scores on both tests that are administered to students at the same time. The condition of "at the same time" is difficult to meet under the current MSPAP/CTBS5 testing plan. Practically, one set of matched-sample MSPAP/CTBS4 data was collected (called CTMS, refer to Figure 3) from the "same 1995 school year". This data included 6824 students' scores on both measures, in which three CTBS/4 Fall content scores (CTBS Reading Vocabulary, CTBS Reading Comprehension and CTBS Math Applications),

along with two MSPAP Spring content scores (MSPAP Reading and MSPAP Math) were selected for SEM/MTMM modeling.

2. Data Analysis and Evaluation

Four specific SEM models (Widaman, 1985), as illustrated previously, in the five content area scores, as indicated above, were created for testing the convergent validity and also discriminant validity.

The distribution of the Satorra-Benter scaled chi-square statistics (Satorra & Bentler, 1988) is more closely approximated by the chi-square value than the usual (or unrescaled) chi-square statistic when the assumption of normality for variables is not held. Because the distribution of test scores used in this study was not normally distributed, the Satorra-Benter scaled chi-square statistic (Satorra & Bentler, 1988) was used for testing the data-model fit for each of the four models.

The variance of the error term for each measure was approximated by: Variance times (1-Cronbach's alpha), for example, the error variance of the MSPAP Reading = 2055.657 (Variance of MSPAP Reading) multiplied by (1-0.81), where 0.81 is the Cronbach's alpha for MSPAP Reading. The error-term variances were then constrained while applying the four SEM models. Since there are only 10 (5x4/2) unique elements for the five-variable MTMM matrix, the degree of freedom could be negative if more than 10 SEM/MTMM parameters are estimated. Hence, fixing the error variances will make the degree of freedom for each SEM model available to test the appropriateness of the specified model.

IV. Results and Discussions

A. Evaluating Longitudinal True-score Associations

CTBS

The analyses of the longitudinal true-score intercorrelations in the CTBS1 dataset were conducted and presented in Table 1. Similar analyses were conducted for the dataset of CTBS2. The criterion for gauging the existence of construct validity for each content measure was evaluated. For the CTBS1 dataset, this criterion was achieved for Reading, Language, and Mathematics. For example, the longitudinal true-score correlation of Reading was .675, which was larger than its longitudinal true-score correlations with other content areas (e.g., .636 for

Reading in 1999 with Language in 1997; .613 for Reading in 1997 with Math in 1999). However, this criterion was not met for Language Mechanics and Mathematics Computations. For example, the longitudinal true-score correlation of Math Computations was .582, which was smaller than its correlations with other content areas (e.g., .630 for Math Computations in 1997 with Math in 1999) . This criterion was also achieved for Reading and Mathematics for the CTBS2 dataset (see Table 2).

To summarize, based on the evaluation of the longitudinal true-score correlations for the multiple-choice CTBS/5 program, the subtests of Reading and Mathematics met the criterion. This finding implies that these two subtests more closely measured their corresponding underlying constructs. This provides evidence of construct validity for these two subtests. More specifically, the measures of the test items of Reading and Mathematics have been constructed so as to measure what they were designed to measure. In contrast, the test items on the other contents (or subjects) may not be clearly designed to assess what they are supposed to measure.

Table 1: Longitudinal True-score Intercorrelations for the Matched-sample CTBS1 dataset: Grade 2 CTBS in 1997 with Grade 4 CTBS in 1999 (N=6841)

		Grade 2 CTBS data in 1997				
	Contents	Reading	Language	Language Mechanics	Math	Math Computations
Grade 4 CTBS Data in 1999	Reading	**.675**	.636	.533	.566	.500
	Language	.643	**.645**	.543	.572	.516
	Language Mechanics	.562	.548	**.537**	.548	.517
	Math	.613	.627	.538	**.683**	.630
	Math Computations	.478	.502	.432	.542	**.582**

Table 2: Longitudinal True-score Intercorrelations for the Matched-sample CTBS2 dataset: Grade 2 CTBS in 1998 with Grade 4 CTBS in 2000 (N=6899)

	Contents	Grade 2 CTBS data in 1998				
		Reading	Language	Language Mechanics	Math	Math Computations
Grade 4 CTBS Data in 2000	Reading	**.668**	.608	.528	.562	.495
	Language	.657	**.632**	.557	.586	.529
	Language Mechanics	.554	.593	**.552**	.536	.509
	Math	.605	.618	.540	**.670**	.614
	Math Computations	.483	.502	.473	.550	**.580**

MSPAP

The analyses of the longitudinal true-score intercorrelations in the dataset of MSPAP1, as well as in the dataset of MSPAP2, were conducted and presented in Tables 3 and 4. The criterion indicated above was achieved for Language and Social Studies, but not for Reading, Writing, Mathematics, and Science for the MSPAP1 dataset. For the MSPAP2 dataset, only the content area of Language met the requirement of this criterion. MSPAP1 dataset had one more subset that met the requirement for the construct validity than the MSPAP2 dataset. This phenomena is not unusual in the examination of the real test data. That is why researchers should examine as much more real test data as possible before drawing an accurate conclusion.

To summarize, based on the results of evaluating the longitudinal true-score correlations for the performance-based MSPAP testing program, the subtests of Language met the criterion in the two sets of data being examined. This finding implies that the measure of the Language subtest was more closely related to its corresponding underlying construct. The evidence of construct validity for this subtest was to some extent established.

Table 3: Longitudinal True-score Intercorrelations for the Matched-sample MSPAP1 Dataset: Grade 3 MSPAP in 1997 with Grade 5 MSPAP in 1999 (N = 6236)

	Contents	Grade 3 MSPAP data in 1997					
		Reading	Writing	Language	Math	Science	Social Studies
Grade 5 MSPAP Data in 1999	Reading	**.604**	.578	.544	.565	.612	.617
	Writing	.588	**.639**	.594	.547	.586	.604
	Language	.676	.727	**.741**	.631	.667	.696
	Math	.624	.625	.582	**.683**	.690	.677
	Science	.649	.661	.599	.650	**.693**	.710
	Social Studies	.654	.657	.585	.633	.684	**.712**

Table 4: Longitudinal True-score Intercorrelations for the Matched-subject MSPAP2 Dataset: Grade 4 MSPAP in 1998 with Grade 5 MSPAP in 2000 (N = 6547)

	Contents	Grade 3 MSPAP data in 1998					
		Reading	Writing	Language	Math	Science	Social Studies
Grade 5 MSPAP Data in 2000	Reading	**.585**	.577	.542	.522	.599	.600
	Writing	.579	**.611**	.584	.519	.591	.592
	Language	.642	.692	**.723**	.582	.663	.657
	Math	.610	.608	.586	**.647**	.668	.627
	Science	.623	.621	.581	.594	**.663**	.642
	Social Studies	.623	.605	.572	.585	.658	**.649**

B. The SEM Approach to Multitrait-multimethod Associations of CTBS/4 and MSPAP

The Satorra-Bentler scaled chi-square value for the hypothesized model M1 (Correlated Traits and Correlated Methods) is 243.778 (see Table 5). Hu and Bentler (1999) recommended two joint criteria to retain a model, such as (CFI >= .96 and SRMR <= .10) or (RMSE <= .06 and SRMR <= .10). This model is presumed appropriate according to one of the above two joint criteria. The standardized path coefficients of Model M1 are given in Figure 1. For example, the standardized path coefficients from the latent trait of READING to the observed variable of MSPAP Reading is .785.

The chi-square value and the goodness-of-fit statistics for the Model M2 (No Traits and Correlated Methods) are presented in Table 5. As indicated by the fit indices, the goodness of fit for Model M2 was poor.

Table 5
Hypothesis Tests and Fit Indices for Models from M1 to M4 (N=6824) in 1995

Model	Satorra-Bentler Scaled Chi-square	df	P	CFI	SRMR	RMSE
M1: Correlated Traits & Correlated Methods	243.778	3	.001	.960	.028	.168
M2: No Traits & Correlated Methods	1534.543	9	.001	.580	.081	.312
M3: Perfectly Correlated Traits & Correlated Methods	439.997	4	.001	.900	.040	.229
M4: Correlated Traits & Perfectly Correlated Method	430.286	4	.001	.902	.046	.227

Model Comparison	Difference in			
	Chi-square	df	P	ΔCFI
Test of Convergent Validity (M1 vs. M2)	1290.765	6	.001	.379
Test of Discriminant-Validity: Traits (M1 vs. M3)	196.219	1	.001	.059
Test of Discriminant Validity: Methods (M1 vs. M4)	186.508	1	.001	.057

Testing Convergent Validity

The evidence of convergent validity was tested by comparing Model M1 with Model 2. A significant difference in chi-square values (see Table 5) between the two models was found and the difference in practical fit (CFI =.379) was substantial. Therefore, the convergent validity was concluded. The convergent validity explains the extent to which two different assessments (e.g., MSPAP and CTBS) of the same trait are correlated. The multitrait-multimethod true-score intercorrelation matrix is presented in Table 6. The true-score correlations between MSPAP Reading with CTBS Vocabulary, as well as with CTBS Reading Comprehension, were .67 and .64. The correlation between MSPAP Math and CTBS Math Applications was .67. Higher correlations are another evidence of convergent validity.

Table 6:
Multitrait-multimethod True-score Correlation Matrix, MSPAP and CTBS/4
(N=6824)

Contents	MSPAP Reading	CTBS Reading Vocabulary	CTBS Reading Comprehension	MSPAP Math	CTBS Math Application
MSPAP Reading	(81)*				
CTBS Reading Vocabulary	.67	(71)			
CTBS Reading Comprehension	.64	.84	(79)		
MSPAP Math	.72	.59	.59	(89)	
CTBS Math Application	.67	.82	.78	.67	(72)

* Values in parenthesis are Cronbach's Alpha coefficients

Testing Trait Discriminant Validity

The chi-square value and the goodness-of-fit statistics for Model M3 (Perfectly Correlated Traits and Correlated Methods) are presented in Table 5. We see that the fit of this model is fairly good, albeit slightly less fitting than for Model M1. In testing for evidence of trait discriminant validity, a significant difference in chi-square values between Model 1 and Model 3

was found (see Table 5). However, the difference in practical fit (CFI =.059) was not substantial. It seems likely that there was weak evidence to support trait discriminat validity.

The trait discriminant validity is the extent to which two different traits measured by different assessment methods are correlated. Lower correlations give evidence of discriminant validity of trait; in contrast, higher correlations connote evidence of poor discriminant validity of the trait. The true-score correlation between MSPAP Reading and CTBS Math Applications was .67. The true-score correlation between MSPAP Math and CTBS Vocabulary (or CTBS Reading Comprehension) was .59.

When we focus on measuring the Math trait, it is unavoidable that the Reading trait is also involved in the process of measuring the Math trait. This factor will cause the poor trait discriminant validity to occur, as happened here. We might improve the trait discriminant validity for the assessments of both Reading and Math only if the Math test items are rewritten with clear wording but with as little reading skill requirement as possible.

Testing Method Discriminant Validity

The chi-square value and the goodness-of-fit statistics for the Model M4 (Correlated Traits and Perfectly Correlated Methods) are presented in Table 5. The fit of this model is almost as good as Model M3, albeit slightly less fitting than for Model M1. In testing for evidence of method discriminant validity, we applied the same logic as noted earlier. A significant difference in chi-square values between these two models of M1 and M4 was found (see Table 5). However, the difference in practical fit (CFI =.057) was not substantial. We therefore conclude weak evidence of assessment method discriminat validity.

The method discriminant validity represents the extent to which the two different traits (e.g., Reading and Math) measured by the same assessment method are correlated. The correlation between MSPAP Reading and MSPAP Math was .72. The correlations of CTBS Math with CTBS Vocabulary and CTBS Math with CTBS Reading Comprehension were .82 and .78, respectively. Higher correlations imply poor discriminant validity of assessment method. Part of the reason for these results is due to the fact that CTBS/4 and MSPAP were administered at two different time periods. Hence, caution should be exercised when interpreting these results.

Comparison of Variance Components

A more specific comparison of trait- and method-related variance can be ascertained by examining the variance components on each measure accounted for by the latent trait, method and error factors. For example, breaking down the variance of MSPAP Reading is illustrated as follows. The trait component of MSPAP Reading in Table 7 equals 0.62 that was computed by squaring the path coefficient of 0.785, from the factor of READING to the observed variable of MSPAP Reading (see Figure 1). The component of assessment method equals 0.21 that was computed by squaring the path coefficient of 0.458, from the factor of MSPAP to the observed variable of MSPAP Reading (see Figure 1). Finally, the component of measurement error equals .19 that was computed by squaring the error-term path coefficient of .435.

Using a similar approach, the variance of each of the five observed measures was subdivided and the results summarized in Table 7. Further scrutiny of the variance components presented in Table 7 reveals the likelihood of method effects for attenuating the trait effects. For instance, the Method effect might play a substantive role in accounting for the variance of MSPAP Math. The trait effect of the CTBS/4 reading comprehension was also attenuated by the multiple-choice assessment method. The results from the variance component analysis seem to imply that either the performance-based assessment or the multiple-choice assessment can attenuate the trait effects.

Comparing the results of the decomposed variance components from this study with those from Li, et al.'s study (1999, refer to the values in the parenthesis in Table 7), both results were very similar. This implies that the model used in Li et al.'s study is likely to fit similar MSPAP/CTBS4 data, for instance, the data used in this study.

Table 7:
Variance Components due to Trait, Assessment Method and Measurement Error for Model M1

Content	Trait		Assessment Method		Measurement Error
MSPAP Reading	READING	.62 (.58)*	Performance	.21 (.24)	.19 (.18)
CTBS Reading Vocabulary		.24 (.23)	Multiple-choice	.47 (.48)	.29 (.29)
CTBS Reading Comprehension		.16 (.18)	Multiple-choice	.63 (.61)	.21 (.21)
MSPAP Math	MATH	.18 (.19)	Performance	.71 (.70)	.11 (.11)
CTBS Math Applications		.43 (.47)	Multiple-choice	.29 (.25)	.28 (.28)

* Values in parenthesis came from Li et al.'s study (1999).

V. Summary and Conclusion

The primary concern of this study was to examine the construct validity for the multiple-choice (CTBS) and the performance-based (MSPAP) testing programs by means of evaluating the longitudinal test datasets in one school district.

The following criterion was evaluated: the true-score correlation between two time-period measures of the same content area is higher than its longitudinal true-score correlations with other content areas. This criterion was achieved for the content areas of Reading and Mathematics in two CTBS datasets and met for the content area of Language in two longitudinal MSPAP datasets.

These results prove the proposition that the criterion used for gauging the construct validity is not easily achieved for all content measures. Reasons for the multiple-choice CTBS testing program could be: (a) CTBS/5 Survey is a relatively unreliable test (e.g., = .79 for Grade 2 Language Mechanics subtest, refer to the CTBS technical report) due to the small number of test items (20 items), (b) the constructs for some of the multiple subtests may not be clearly defined or indistinguishable (e.g., Math Concept and Math Computations), and (c) the two factors combined. The above reasons can also be applied to the performance-based MSPAP

testing program. In addition, the fact that MSPAP 's test tasks (or items) were integrated across content areas is another factor affecting the likelihood of achieving that criterion..

The finding from the evaluation of the longitudinal true-score associations might threaten the construct validity of MSPAP and CTBS and brings up the broad question of whether some content-area scores obtained on MSPAP or CTBS reflect the efficacy of the instructional programs in schools, school districts, and the State. It is not appropriate for us to prejudge this issue because of several questions associated with this finding. For example, (1) How this result can be generalized to the test data of other school districts or the whole State? (2) How the criterion used to gauge the construct validity can be achieved when the period of time between two measures is shorter (e.g., only one year period)? and (3) Is this criterion, in practice, too difficult to hold for the testing program with more than five content area measures (e.g., CTBS has five and MSPAP has six content areas measures)? These questions will require clarification at some future time.

The results from SEM to the multitrait-multimethod data, where the Reading and Mathematics traits were assessed by MSPAP and the old version of CTBS/4, suggest that although convergent validity for these two measures existed, the evidence to support the discriminant validity for these two measures was weak. Also, the results from the variance component analysis seem to imply that either the performance-based assessment or the multiple-choice assessment can attenuate the trait effects. More studies need to be done for better understanding the construct validity for each multiple-content testing program in order to ensure the data delivered by the operationally testing program is valid.

Reference:

Bentler P. M. (1995). EQS Structural equations program manual. Encino, CA: Multivariate Software, Inc.

Byrne, B. M. & Bazana, P. G. (1996). Investigating the measurement of social and academic competencies for early/late preadolescents and adolescents: A Multitrait-multimethod analysis. Applied Measurement in Education, 2, 113-132.

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait- multimethod matrix. Psychological Bulletin, 56, 81-105.

Crocker, L. & Algina, J. (1986). Introduction to classical & modern test theory. Orlando, Florida: Holt, Rinehart and Winston, Inc.

CTBS/McGraw-Hill (1997). Teacher's guide to TerraNova. Monterey, CA. McGraw-Hill Companies, Inc.

CTBS/McGraw-Hill (2001). Technical report. Monterey, CA. McGraw-Hill Companies, Inc.

Hu, L.& Bentler, P. M. (1999). Cutoff criterion for fit indexes in covariance structure analysis: Conventional criterion versus new alternatives. Structural Equation Modeling: A multidisciplinary Journal, 6, 1-55.

Kolen, M. J. & Brennan, R. L. (1995). Test equating: Methods and practices. New York: Springer-Verlag.

Li, Y. H. , Ford, V. & Tompkins, L. J. (1999, April). The construct validity of a performance-based assessment program. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Linn, R. L. (1995). High-stakes uses of performance-based assessments: Rationale, examples, and problems of comparability. In T. Oakland & R. K. Hambleton (Ed.), International perspectives on academic assessment (pp. 49-73).Norwell, MA. Kluwer Academic Publishers.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. New Jersey: Lawrence Erlbaum Associates, Inc.

Maryland State Department of Education. (1998). Technical report: 1998 Maryland School Performance Assessment Program. Baltimore: Author.

Nunnally, J. C. & Bernstein, I. H. (1994). Psychometric theory. New York: McGraw-Hill, Inc.

Schatz, C. J. (1998, November). Convergent-discriminant validity evidence for the MSPAP

Satorra, A. & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), Latent variables analysis: Applications for development research (pp. 399-419). Thousand Oaks, CA: Sage.

Schatz, C. J. (1998, November). Convergent-discriminant validity evidence for the MSPAP Reading and Math scores. Paper presented at the annual meeting of the Maryland Assessment Group, Ocean City, MD.

Schmitt, N. & Stults, D. M. (1986). Methodology review: Analysis of multitrait-multimethod matrices. Applied Psychological Measurement, 10, 1-22.

Widaman, K. F. (1985). Hierarchically tested covariance structure models for multitrait-multimethod data. Applied Psychological Measurement, 9, 1-26.

Yen, W. M. & Ferrara, S. (1997). The Maryland school performance assessment program: Performance assessment with psychometric quality suitable for high stake usage. Educational and Psychological Measurement, 57, 60-84.

Yen, W. M. & Ferrara, S. (1997). The technical quality of performance assessments: Standard errors of percents of pupils reading standards. Educational Measurement: Issues and Practice, Fall, 5-15.